International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)
ISSN(P): 2249-6890; ISSN(E): 2249-8001
Vol. 9, Issue 2, Apr 2019, 725-736

A COST-EFFECTIVE TECHNIQUE TO AVOID COMMUNICATION AND COMPUTATION OVERHEAD IN VEHICLE INSURANCE DATABASE FOR ONLINE RECORD MONITORING

ESWARAIAH THUMMULA[1], RAVI KANT YADAV[2] & AROKIARAJ DAVID[3]

[1]*Research Scholar, Department of Computer Science, VBS Purvanchal University, Janpur, Uttar Pradesh, India*

[2]*Professor, Department of Computer Science, VBS Purvanchal University, Janpur, Uttar Pradesh, India*

[3]*Associate Professor, Department of Management Studies, Aurora Post Graduate College, Hyderabad, Telangana, India*

ABSTRACT

The study focuses on the information from heterogeneous data sources in online vehicle insurance record linkage. It resolves several types of heterogeneity problems that arise, when the same real-world entity type is represented using different identifiers in multiple data sources. Statistical vehicle insurance data record linkage techniques are used to resolve the problem. These techniques are used in vehicle insurance online record linkage, which creates good communication bottleneck in the distributed platform. It was projected by the Identical Tree and Decision Tree. And, to reduce the communication barriers, it is significantly pointed while matching with the decisions that are guaranteed to be same as those obtained victimizations with the standard linkage technique. The sequential record linkages along with matching tree using tree-based linkage techniques are used to improvised the accuracy of record linkage technique and reduce the communication overhead in Vehicle insurance Database.

KEYWORDS: Online Record Linkage, Network, Data Warehouse, Data Mining, Data Integration, Data Accessing

Received: Feb 11, 2019; **Accepted:** Mar 01, 2019; **Published:** Mar 27, 2019; **Paper Id.:** IJMPERDAPR201972

Original Article

INTRODUCTION

Overview of Data Warehouse

The warehouse is uploaded from the operational system. The knowledge might tolerate associate with the operational data store to add operations before it is utilized in the data warehouse for coverage [12].

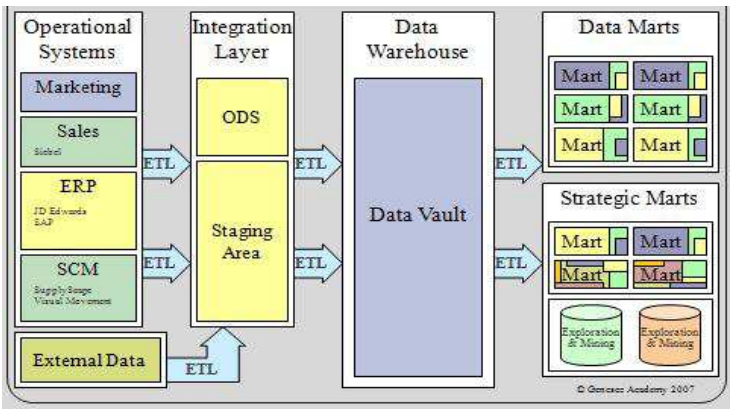

Figure 1: Data Warehouse Architecture

A data warehouse maintains its functions into three layers are staging, integration, and access. Staging is employed to store data to be used by the developers. The integration layer is employed to integrate knowledge and to possess a grade of abstraction from the users. The access layer is obtaining knowledge out for the users [5]. It classified the knowledge. Info warehouse focuses on data storage.

The main supply of the information is cleaned, transformed, cataloged and made available for use by managers and other business professionals for data mining, online analytical processing, market research, and decision support. However, it suggests that to retrieve and analyze knowledge, extract, transform, load data, and to manage the data dictionary are also considered as essential components of a data warehousing system. Many references related to knowledge reposition were used. Thus, it associated with a distended definition for knowledge repositing includes business intelligence tools, tools to extract, remodel and cargo knowledge into the repository, and tools to manage and retrieve information.

Application Layers in Data Warehouse

The layers of applications are present in a Data Warehouse design. It gives an idea about what it deals with. Note, this is just a basic representation of how most standard data warehouses are implemented [12]. There are deviations from what is discussed here, based on the business need analysis and their decisions.

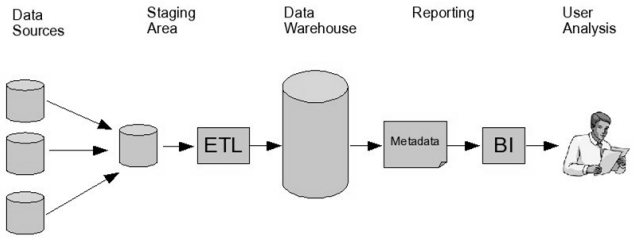

Figure 2: Data Warehouse Application Layers

Data Mining

This falls under the Business Intelligence section, which acts of identifying patterns in the gathered data. The term actually digs into data and tries out various permutations to identify an emerging pattern that could be useful to make an improvised decision. For instance, a pattern could emerge that states that a specific product or brand sales more on the internet rather than on the market shelf in a certain geographical location and could result in tax savings [3].

The Drawback of Existing Systems

- The Vehicle insurance databases are distributed heterogeneity in nature, and not possible to create a central data repository or warehouse where pre-computed linkage results can be stored.
- If the insurance databases span several organizations, the problem with ownership and cost allocation related to the warehouse. Even, if the warehouse could be developed, it would be difficult to up-to-date the data.
- As update occurs at the operational level, the vehicle insurance data linkage results would become stale, if they are not updated immediately.
- The participating sites allow controlled sharing of portions of their vehicle insurance databases using standard database queries, but they do not allow the processing of scripts, stored procedures, or other application programs

from another organization. The problem can be sort-out by both technological abilities, and also management and control [2].

PROPOSED SYSTEM

Here, we proposed a system to draw upon the research in the area of sequential information acquisition to provide an efficient solution to the online, distributed vehicle insurance data record linkage problem. The main benefit of the sequential approach is that, it does not search all the attributes of all the remote records are brought to the local site attributes are bring into one at a time. In sequential approach, as possible matches, the same set of records as the traditional full-information case were brought to the local site (where all the attributes of all the remote records are downloaded).

Advantages

- The sequential approach decides on the next "best" attribute to acquire, based upon the comparison results of the previously acquired attributes data.
- The communication overhead as low as possible. The partitioning itself can be done in one of two possible ways (sequential and concurrent).
- SIA queries would involve selecting vehicle insurance records by comparing values of attributes for which, a secondary index may not exist.

Project Goal: A distinct advantage of the tree-based sequential record linkage is that the matching tree can be pre-compiled and stored, there, saving computational overhead at the time of answering a linkage query.

SEQUENTIAL APPROACH RECORD LINKAGE AND MATCHING TREE

In this section, it describes the sequential information acquisition to provide an efficient solution to the online, distributed vehicle insurance record linkage problem. The main benefit of the sequential approach is that, unlike the traditional full-information case, not all the attributes of remote records are brought to the local site; instead, attributes are brought one at a time. After acquiring an attribute, the matching probability is revised based on the realization of the attribute, and a decision is made whether to acquire more attributes or not.

The sequential approach decides on the next "best" attribute to acquire, based upon the comparison of results to the previously acquired attributes [15]. The acquisition of attributes can be expressed in the form of a matching tree as shown below. There are two basic principles used in the induction of a matching tree.

- Input selection
- Stopping

Before we describe these two principles, we have to clarify an important point are subsequent numerical analysis. It makes the common assumption of the conditional independence that reduces the overall computational burden.

Input Selection: Assume that 'i' is at some node of the tree and trying to decide how to branch from there. The set of attributes that have already been acquired; the possibility that is not excluded [12]. The matching probability as revised by the attributes can be written. At this point, interest in finding the next best attribute to be acquired from the set of remaining attributes.

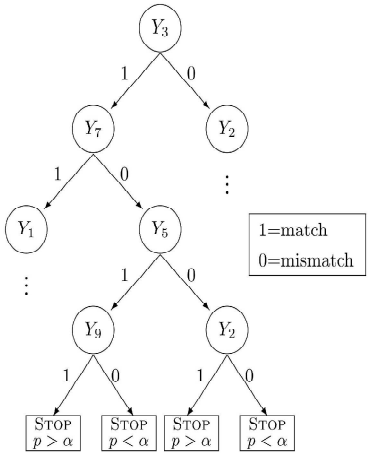

Figure 3: A Sample Tree Showing Attributes Acquisition Order

Stopping: Now, consider the issue when to stop or expanding the matching tree. The stopping decision is made when no realization of the remaining attributes can sufficiently revise the current matching probability, so that the matching decision changes. To that end, it finds the upper and lower bounds of the eventual matching probability.

TREE-BASED LINKAGE TECHNIQUES

The advance efficient online vehicle insurance record linkage techniques based on the matching tree induced. The overall vehicle insurance linkage process is summarized below. In the first two stages, the process is performed offline by using the training data. The matching tree has been built and the online linkage is professed in the final stage.

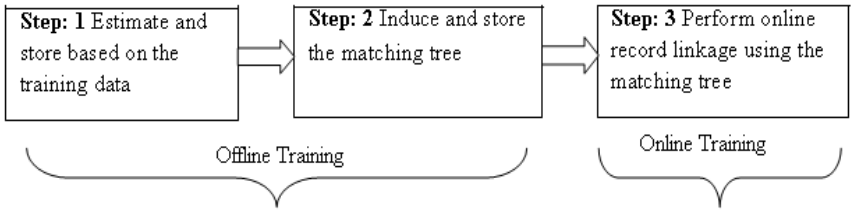

Figure 4: The Overall Process of Online Tree-Based Linkage

It can characterize the different techniques that can be employed in the last step. Recall that, given a local inquiry record, the ultimate goal of any linkage technique is to identify and fetch all the records from the remote site that has matching probability [4]. In other words, one needs to partition the set of remote records into two subsets,

- Relevant records that have a matching probability
- Irrelevant records that have a matching probability

The aim is to develop techniques that would achieve a better objective, while keeping the communication overhead as low as possible, the partitioning has been done in two possible ways are sequential and concurrent

SEQUENTIAL ATTRIBUTE ACQUISITION (SAA)

Here, 'I' acquire attribute from the remote records in a sequential fashion. Consider the matching tree working with this tree, 'i' would first acquire attribute for all the remote records. When the value is compared with the local inquiry record, 'i' would get either a match or a mismatch.

A vehicle insurance record identification scheme common to both the sites must be established. This can be easily done by using a candidate key from the remote database. In this scheme, during the first transfer, 'I' acquire the identifiers for all the remote records and use these identifiers to specify any desired partition of the set of remote records. The total communication overhead of the SAA technique is composed of three elements.

- The transfer of attribute values from the remote to the local site.
- It transfers all the identifiers between the local sites and the remote.
- It also transfers these vehicle insurance data records matching probability. It is possible to estimate that expected size of each of these three overheads from the matching tree.

SEQUENTIAL IDENTIFIER ACQUISITION (SIA)

Sequential identifier acquisition is a minor variation of sequential attribute acquisition. It can provide significant savings in terms of the communication overhead. It also better performance lies, in fact, nonkey attributes stored in a database are often much larger than an identifier. If the attribute transfer could be replaced by identifier transfer, the total communication may get reduced. Therefore, in this approach, it is not possible to transfer the attribute values from the remote site.

The local inquiry record, and ask the remote database to send the identifiers from only that subset of which matches. Proceeding in this way, 'i' can eventually find the identifiers of all the remote records with a matching probability greater. In this case, there are three types of communication overheads: 1) attribute overhead, 2) identifier overhead, and 3) included record overhead. In order to obtain the total attribute, overhead is noted that the attribute value of the inquiry record at a node x must be sent as long as there is a single remote record that visits.

CONCURRENT ATTRIBUTE ACQUISITION (CAA)

The actual performance of the above approaches, implement and tested them on real-world and synthetic datasets. Before describing the implementation and discuss the results, two aspects of the numerical study should be discussed.

The expected communication overhead for the sequential approach (normalized by the size of the remote database) can be calculated exactly based on the matching tree. Hence, they are a need not resort to simulation (using actual data sets) to estimate the expected communication overhead. The communication overhead for each value of can then is calculated. It focuses on the efficiency or performance of the approaches in terms of reducing communication overhead.

DISTANCE BASED CLUSTERING

Data cluster is a method by which, a large set of data is grouped into clusters of smaller sets of similar data. It has assigned a distance by measuring between the data to partition,

- The distance between objects within the partition (i.e. same cluster) is minimized.
- The distance between objects from different clusters is maximized.

Inputs and Outputs of Database Integration

One of the basic principles of the information approach is that information permits a non-redundant, unified illustration of all knowledge managed in Vehicle Insurance companies. This is true only when methodologies are available to support integration across organizational and application boundaries. More and more organizations are becoming aware of the potential of vehicle insurance database systems and wish to integrated applications with software for the fast retrieval and data update. Even when applications and user groups are structurally disconnected, as in most governmental and large administrative setups, there is something to be gained by having an enterprise-wide view of the data resource.

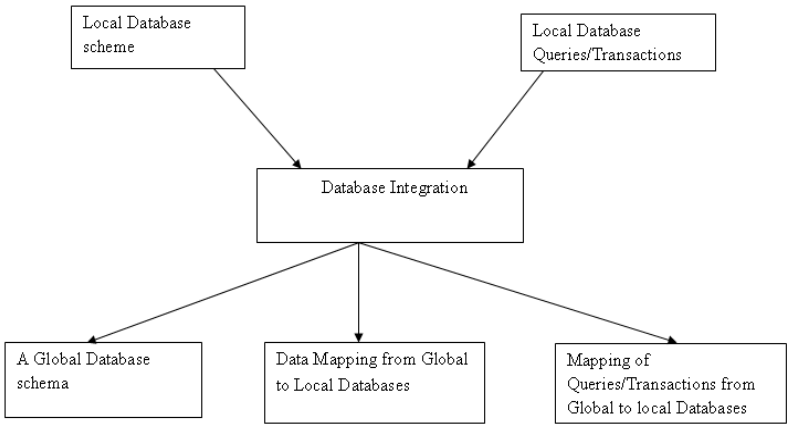

Figure 5: Data Integration

PROCESS DIAGRAM OF A RECORD LINKAGE SYSTEM

Record linkage techniques square measure the accustomed link along records, which relate to a similar entity (e.g. patient or customer) in one or a lot of knowledge sets, wherever a singular symbol isn't accessible [6]. Record linkage is a crucial initial step in several analyses and data processing in medicine and other different sectors. It is used to improve data quality and to assemble longitudinal or different knowledge sets, which might not rather be accessible.

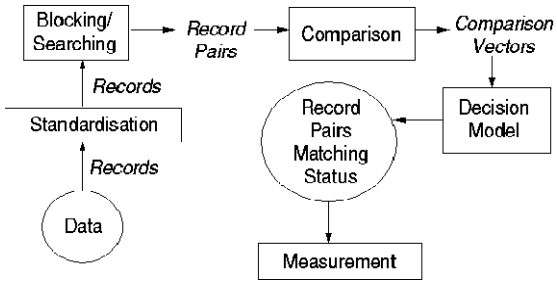

Figure 6: Record Linkage Process

SYSTEM MODULES

Data Accessing: The performance and management of business operations make and implementation of major works. Some plans like Money Back Policies (MBP) provides a medic claim to the vehicle insurance policyholders provided premium due, under the vehicle insurance policies are paid up to the due for survival benefit. In these cases, where the amount payable to vehicle insurance, bill amount to policy holder cheque is released after, calling for the discharge receipt or policy document.

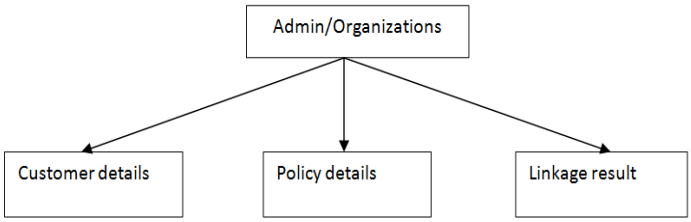

Figure 7: Data Accessing by the Administrator

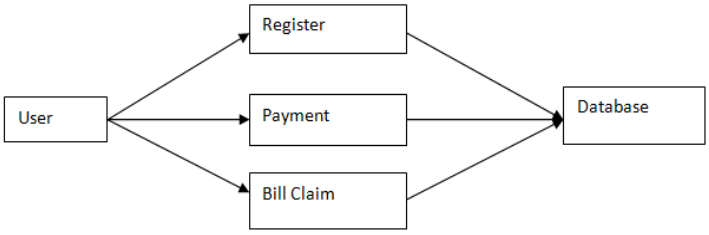

Figure 8: Data Accessing by the User

Data Comparison: There are various types of Vehicle insurance policies. Some of them are considered for human life and health benefits, whereas some focus of the private belongings of someone. The terms and conditions of insurance plans differ from each other. Some of the common insurance policies

Figure 9: Policy Details

In this module, manages keep track of Vehicle insurance policy claims that are raised by the vehicle insurance policy holders. Its priority check is deal with modules of policy payments and policy info modules. This module integrates with on-top of two modules to stay track of the specification like consistency and integrity.

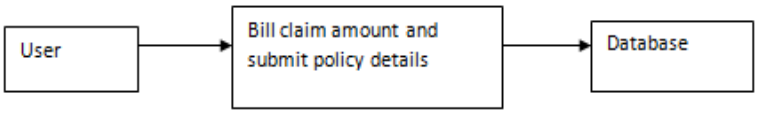

Figure 10: Policy Claims

In the module, manages keep track of the vehicle insurance policy payments by the registered vehicle insurance policy holders. It has interaction to Vehicle policy information module to keep track of the consistency of information from time to time. This module standardizes the security issues that come up on to the system when an authorized person should make his entry into the insurance database. The system manages the information related to the authorized staffs that are entitled to work upon the existing database in a confidential manner.

The Linkage between Records: Record linkage is a vital issue in heterogeneous information systems, wherever the vehicle insurance records representing a real-world entity sort square measure known victimization that are totally different identified in numerous databases. In the absence of a common identifier, the matching probability is computed based on common attribute values. This needs those common attribute values of all the remote records to be transferred to the native web site and avoid communication overhead.

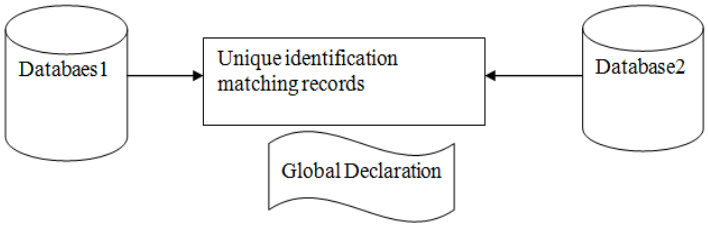

Figure 11: Record Linkage between Two Databases

Requirements are the basic needs or constraints, which are required to develop a system. These requirements can be collected while designing the system. Here, are two main classifications for the requirements; they are user requirements and system requirements. The following requirements are to be discussed below.

Design and Implementation: The design phase generally consists of following diagrams such as sequence diagram, a flow diagram, a process flow diagram, a collaboration diagram. Each diagram explains the notion of our proposed system.

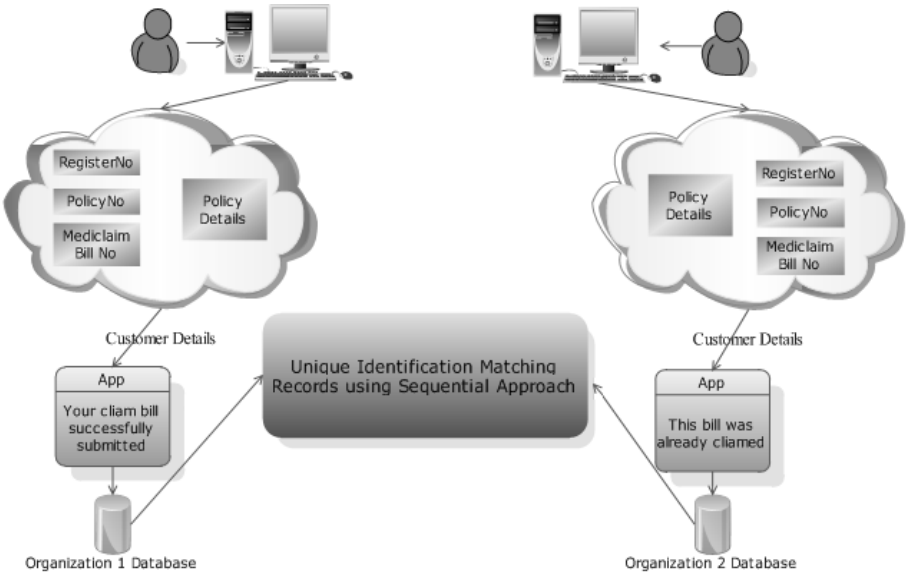

Figure 12: System Architecture

The systems architecture is responsible for interfacing with the user(s) and sponsor(s) and other stakeholders are determining their (evolving) needs. Generating the best level of system needs, based on the user's needs and other constraints such as cost and schedule [12]. Performing cost-benefit analysis is to see whether or not needs square measure best is met by manual, software, or hardware functions by making maximum use of commercial off-the-shelf or already developed components.

SEQUENCE DIAGRAM

A sequence diagram is Unified Modeling Language (UML) that could be a reasonable interaction diagram that shows how processes operate with each other and in what order. It is a construct of a Message Sequence Chart (MSC). A sequence diagram shows the object of interactions, which is organized in time sequence. It depicts the objects and categories concerned within the situation, therefore the sequence of messages changed between the objects required to hold out the Practical situation.

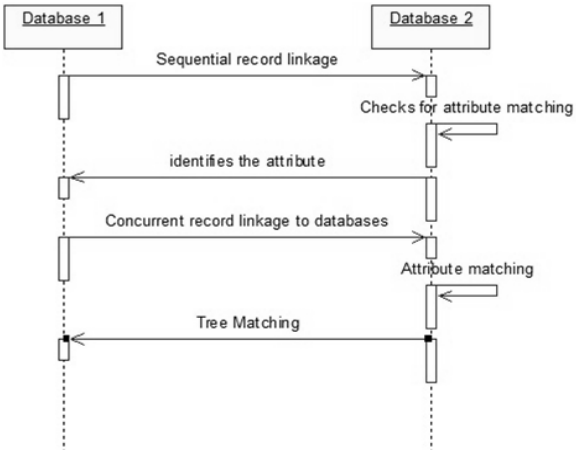

Figure 13: Sequence Diagram

STATE DIAGRAM

The state diagram within the UML is actually a Harel State Chart (HSC) with standardized notation, which might describe several systems, from PC programs to business processes.

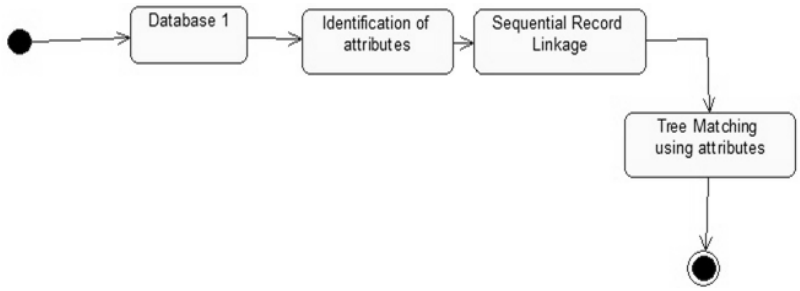

Figure 14: State Diagram

USE CASE DIAGRAM

A use case diagram within the UML may be a style of activity diagram outlined by and created from a use-case analysis. The purpose is to provide a graphical summary of the operations provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between the cases.

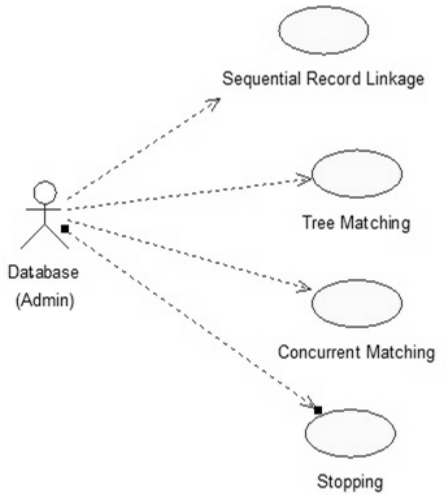

Figure 15: Use Case Diagram

Implementation: Using Microsoft visual studio 2010 and SQL server 2008, create the modules for access the Vehicle insurance data of the client, with the back-end connection. This makes all the data to secure by the admin. Coding is programmed in C# which access and perform modifications in the multiple databases connected [12]. Common attributes collection is the major and important task of this work. Identifying common attributes in the same column of the multiple vehicle insurance databases is implemented with overall concentrate. Accessing the column of the vehicle insurance database in a secure and efficient manner is handled by the built-in database connectivity manager. The specific data bundles are collected in the categorical order that is in columns in the vehicle insurance database table. The handling of multiple databases also a crucial part of this implementation. If the resulting data set is not upto the satisfactory level, the alternate data source in the same database can be chosen by the user [9]. The finding of the best fit common attribute list is also implemented by some basic preprocessing based on the rules given by the user and by the metadata understandings.

Testing: After successful completion of the coding, code review was done with the objective of identifying and correcting deviations from standards, identifying and fixing logical bugs and fall through and recording code walkthrough findings [12]. The programs were checked, and therefore the code structure was created clear. The variable names were meaningful [8]. It follows certain naming conventions, which makes the program readable. Variable names are prefixed with their scope and data type. Check-out for the correct scopes for various functions. All possible explanations for the code were given as comments [12].

- Sufficient labels and comments are included in the code as the description of it, the benefit of the developer and other programmers who might examine it later.
- Checking out the connectivity of the vehicle insurance database.
- Code optimization was carried out.

Testing Cases: Usual login test cases are applied to the login module testing. Such cases include a wrong password, empty field etc. After successful login in the register page, vehicle insurance customer data are entered. These data have to check before registering to prevent unwanted registrations. These test cases have a variety of input data for each case to be checked in the stage of testing.

Admin and normal users have to be distinguished and the level of access for the particular level of users must be checked based on the requirement proposed. Creation of the agents and management of the agents and the policy under the agents registered are checked of integrity, which has to be maintained. Vehicle insurance Policy maintenance has another set of test cases, which has to check the changes in policy is reflecting well in the database. Other functional based requirements are checked in the testing. After logout, all the session log outs has to be checked for the particular user. The logs are used to find the bugs and faults occur during the user access.

CONCLUSIONS

The efficient techniques to facilitate record linkage decisions in a distributed online setting are concluded. Record linkage is a vital issue in heterogeneous information vehicle insurance database systems. Wherever the records represent, an equivalent real-world entity are victimized using different identifiers in numerous databases. In the absence of a common identifier, it is often difficult to find records in a remote database, which are similar to a local inquiry record.

Traditional record linkage uses a probability-based model to spot the closeness between records. The matching likelihood is computed by common attribute values. It needs those common attribute values of all the remote records to be transferred to the Local website. The communication overhead is considerably massive for associate operation. The matching must be performed in a manner, so that no local record is paired with the remote record and vice versa.

SCOPE FOR FURTHER RESEARCH

Another avenue for future research is to perform an explicit cost-benefit trade-off between error cost and communication overhead. In this study, it reduces the communication overhead significantly while keeping the error cost at the level of traditional techniques. It may, however, be possible to further reduce the communication overhead at the expense of incurring higher costs of vehicle insurance record linkage errors. One could also apply sequential decision-making techniques to the vehicle insurance data record linkage problem, using non-probabilistic similarity measures such as the distance-based measures by using the cluster method. This may be useful in situations, where the training data (to estimate the probabilities) are not readily available.

REFERENCES

1. *Arokiaraj, D. (2015). A Study on Environmental Responsibility of the Stakeholders of Auto Industry in Chennai.*
2. *Arokiaraj, D., R. Srivel, M., & R. Arthie Reena, M. (2018). The Eco-Driving Behaviour: A Strategic Way to Control Tailpipe Emission. International Journal of Engineering & Technology, 7(3.3), 21-25.*
3. *Bushan D. Sudhakar, Kattepogu Nagarjuna & D. Arokiaraj (2017), "Marketing Assistance and Digital Branding- An Insight for Technology Up-gradation for MSME's", International Journal of Management Studies & Research, Vol. No.5, e-ISSN: 2455-1562; Print-ISSN: 2455-6092, ISBN: 978-93-86171-56-6. P415-421*
4. *D. Arokiaraj & R. Srivel, (2017) "The direness of cultivable land spotted on agricultural: A special reference to rice production in South India" National Monthly Refereed Journal of Research in Commerce & Management, ISSN-2277-1166, Vol. 06, issue 09 pp. 55-59.*
5. *D. Dey, S. Sarkar, and P. De, "A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases," IEEE Trans. Knowledge and Data Eng., vol. 14, no. 3, pp. 567-582, May/June 2002.*
6. *D. Dey, S. Sarkar, and P. De, "A Probabilistic Decision Model for Entity Matching in Heterogeneous Databases," Management Science, vol. 44, no. 10, pp. 1379-1395, 1998.*
7. *D.S. Zingmond, Z. Ye, S. Ettner, and H. Liu, "Linking Hospital Discharge and Death Records-Accuracy and Sources of Bias," Journal Clinical Epidemiology, vol. 57, no. 1, pp. 21-29, 2004.*
8. *David, A., & Banumathi, M. (2014). A Study on Eco–Driving Behaviour of Passenger Car Users in Chennai.*
9. *David, A., & Banumathi, M. (2014). Factors Influencing the Purchase Decision of Passenger Cars in Puducherry.*
10. *Gaikwad, N., Mistry, Y., & Inamdar, K. (2016). Design and Implementation of Energy Efficient Environment Monitoring System.*
11. *I. Baussano, M. Bugiani, D. Gregori, C. Pasqualini, V. Demicheli, and F. Merletti, "Impact of Immigration and HIV Infection on Tuberculosis Incidence in an Area of Low Tuberculosis Prevalence," Epidemiology and Infection, vol. 134, no. 6, pp. 1353-1359, 2006.*
12. *L. Gu and R. Baxter, "Adaptive Filtering for Efficient Record Linkage," Proc. Fourth SIAM Int'l Conf. Data Mining (SDM '04), pp. 22-24, Apr. 2004.*

13. *Montano-Soto, Tatiana., Arellano-García, Evarista., & Camarena-Ojinaga, L. (2014). Genotoxic biomonitoring and exposure to pesticides in women laborers at maneadero valley in Baja California, Mexico. International Journal of Applied and Natural Sciences, 3(2), 89-96.*

14. *P. Pratheepkumar, J. Josephine Sharmila & D. Arokiaraj, (2017) "Towards Mobile Opportunistic in Cloud Computing" Indian Journal of Scientific Research (IJSR), Issue02, vol 17, ISSN: 0976-2876 (Print) ISSN: 2250-0138(Online), 536 – 540.*

15. *S.N. Minton, C. Nanjo, C.A. Knoblock, M. Michalowski, and M. Michelson, "A Heterogeneous Field Matching Method for Record Linkage," Proc. Fifth IEEE International Conf. Data Mining (ICDM '05), pp. 314-321, Nov. 2005.*

16. *Srivel, D. Arokiaraj & Mohammed Imaduddin, 2018 "Controlling & Calibrating Vehicle-Related Issues Using RFID Technology" International Journal of Mechanical and Production Engineering Research and Development (IJMPERD), ISSN (P): 2249-6890; ISSN (E): 2249-8001, Vol. 8, Issue 2, pp. 1125-1132, DOI: 10.24247/ijmperdapr2018130.*

17. *Srivel, R., R. P. Singh, D., & D. Arokiaraj, D. (2018). The FPGA implementation of power on self-test towards combo card. International Journal of Engineering & Technology, 7(3.3), 156-161.*